

Selfish Computation Offloading for Mobile Cloud Computing in Dense Wireless Networks

Sladana Jošilo and György Dán

ACCESS Linnaeus Center, School of Electrical Engineering

KTH, Royal Institute of Technology, Stockholm, Sweden E-mail: {josilo, gyuri}@kth.se

Abstract—Offloading computation to a mobile cloud is a promising solution to augment the computation capabilities of mobile devices. In this paper we consider selfish mobile devices in a dense wireless network, in which individual mobile devices can offload computations via multiple access points (APs) to a mobile cloud so as to minimize their computation costs, and we provide a game theoretical analysis of the problem. We show that in the case of an elastic cloud, all improvement paths are finite, and thus a pure strategy Nash equilibrium exists and can be computed easily. In the case of a non-elastic cloud we show that improvement paths may cycle, yet we show that a pure Nash equilibrium exists and we provide an efficient algorithm for computing one. Furthermore, we provide an upper bound on the price of anarchy (PoA) of the game. We use simulations to evaluate the time complexity of computing Nash equilibria and to provide insights into the PoA under realistic scenarios. Our results show that the equilibrium cost may be close to optimal, and the cost difference is due to too many mobile users offloading simultaneously.

I. INTRODUCTION

Mobile handsets are increasingly used for various computationally intensive applications, including augmented reality, natural language processing, face, gesture and object recognition, and various forms of user profiling for recommendations [1], [2]. Executing such computationally intensive applications on mobile handsets may result in slow response times, and can also be detrimental to battery life, which may limit user acceptance.

Mobile cloud computing has emerged as a promising solution to serve the computational needs of these computationally intensive applications, while potentially relieving the battery of the mobile handsets [3], [4]. In the case of mobile cloud computing the mobile devices offload the computations via a wireless network to a cloud infrastructure, where the computations are performed, and the result is sent back to the mobile handset. While computation offloading to general purpose cloud infrastructures, such as Amazon EC2, may not be able to provide sufficiently low response times for many applications, emerging mobile edge computing resources may provide sufficient computational power close to the network edge to meet all application requirements [5].

Computation offloading to a mobile edge cloud can significantly increase the computational capability of individual mobile handsets, but the response times may suffer when many handsets attempt to offload computations to the cloud simultaneously, on the one hand due to the competition for possibly constrained edge cloud resources, on the other hand due to contention in the wireless access [6], [7]. The problem is even more complex in the case of a dense deployment of access points, e.g., cellular femtocells

or WiFi access points, when each mobile user can choose among several access points to connect to. Good system performance in this case requires the coordination of the offloading choices of the individual mobile handsets, while respecting their individual performance objectives, both in terms of response time and energy consumption.

In this paper we consider the problem of resource allocation for computation offloading by self-interested mobile users to a mobile cloud. The objective of each mobile user is to minimize a linear combination of its response time and its energy consumption for performing a computational task, by choosing whether or not to offload through one of many access points. Clearly, the choice of a mobile user affects the cost of other mobile users. If too many mobile users choose offloading through a particular access point then they will achieve low transmission rate. A low transmission rate would lead to high data transmission time and a corresponding high energy consumption. In order to capture the interactions between the choices of the mobile users, in this paper we formulate the computation offloading problem as a non-cooperative game, and address the existence of self-enforcing resource allocations, i.e., equilibrium allocations, and their computation.

Our contributions in this paper are threefold. First, we show that if the cloud computing resources scale with the number of mobile users then equilibrium allocations always exist, and we provide a simple algorithm for computing an equilibrium. Second, we show that if the cloud computing resources do not scale with the number of mobile users then the same algorithm cannot be used for computing an equilibrium as it may cycle infinitely, but we prove that equilibria exist, and we provide an algorithm with quadratic complexity in the number of mobile users for computing an equilibrium. Finally, we provide a bound on the price of anarchy for both models of cloud resources. We provide numerical results based on extensive simulations to illustrate the computational efficiency of the algorithms and to evaluate the price of anarchy for scenarios of practical interest.

The rest of the paper is organized as follows. We present the system model in Section II. We prove equilibrium existence and computability results for the elastic cloud and non-elastic cloud in Sections III and IV, respectively. We provide a bound on the price of anarchy in Section V and present numerical results in Section VI. Section VII discusses related work and Section VIII concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a mobile cloud computing system that serves a set $\mathcal{K} = \{1, 2, \dots, K\}$ of colocated mobile users (MU). Each MU has a computationally intensive task to perform, and can decide whether to perform the task locally or to offload the computation to a cloud server. The computational task is characterized by the size D_k of the input data (e.g., in bytes), and by the number L_k of the CPU cycles required to perform the computation. To enable a meaningful analysis, we make the common assumption that the set of MUs does not change during computation offloading, i.e., in the order of seconds [4], [8], [9], [10].

A. Communication model

If the MU decides to offload the computation to the cloud server, it has to transmit D_k amount of data pertaining to its task to the cloud server through one of a set of access points (APs) denoted by $\mathcal{I} = \{1, 2, \dots, I\}$. Thus, together with local computing MU k can choose an action from the set $\mathcal{D}_k = \{0, 1, 2, \dots, I\}$, where 0 corresponds to local computing, i.e., no offloading. We denote by $d_k \in \mathcal{D}_k$ the decision of MU k , and refer to it as her strategy. We refer to the collection $\mathbf{d} = (d_k)_{k \in \mathcal{K}}$ as a strategy profile, and we denote by $\mathcal{D} = \times_{k \in \mathcal{K}} \mathcal{D}_k$ the set of all feasible strategy profiles.

We denote by B_i the bandwidth of AP i , and for a strategy profile \mathbf{d} we denote by $n_i(\mathbf{d})$ the number of MUs that use AP i for computation offloading, and by $n(\mathbf{d}) = \sum_{i \in \mathcal{I}} n_i(\mathbf{d})$ the number of MUs that offload. Similarly, for an AP $i \in \mathcal{I}$ we denote by $O_i(\mathbf{d}) = \{k | d_k = i\}$ the set of MUs that offload using AP i , and we define the set of offloaders as $O(\mathbf{d}) = \cup_{i \in \mathcal{I}} O_i(\mathbf{d})$. We consider that the bandwidth B_i of AP i is divided equally among the users that are connecting to it, i.e., the uplink rate $R_k^i(\mathbf{d})$ of MU k is given by

$$R_k^i(\mathbf{d}) = \frac{B_i}{n_i(\mathbf{d})}. \quad (1)$$

The model of equal bandwidth sharing is reasonable if MUs are colocated, or if the APs implement fair uplink bandwidth allocation [11], [12].

The uplink rate $R_k^i(\mathbf{d})$ together with the input data size D_k determines the transmission time $T_{k,i}^{c,off}(\mathbf{d})$ of MU k for offloading via AP i ,

$$T_{k,i}^{c,off}(\mathbf{d}) = \frac{D_k}{R_k^i(\mathbf{d})}. \quad (2)$$

To model the energy consumption of the MUs, we assume that MU k uses a constant transmit power of P_k for sending the data, thus the energy consumption of MU k for offloading the input data of size D_k via AP i is

$$E_{k,i}^c(\mathbf{d}) = \frac{D_k P_k}{R_k^i(\mathbf{d})}. \quad (3)$$

B. Computation model

In what follows we introduce our model of the time and energy consumption of performing the computation locally and in the cloud server.

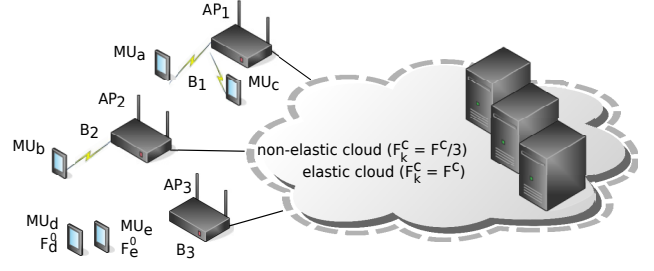


Fig. 1. An example of a mobile cloud computing system

1) *Local computing*: In the case of local computing data need not be transmitted, but the task has to be processed using local computing power. We denote by F_k^0 the computational capability of MU k , and express the time it takes for MU k to perform the computation task $\langle D_k, L_k \rangle$ locally by

$$T_k^0 = \frac{L_k}{F_k^0}. \quad (4)$$

In order to model the energy consumption of local computing we denote by v_k the consumed energy per CPU cycle, thus we obtain

$$E_k^0 = v_k L_k. \quad (5)$$

2) *Cloud computing*: In the case of cloud computing, after the data are transmitted via an AP, processing is done at the cloud server. We denote the computation capability of the cloud by F^c , and by F_k^c the computation capability assigned to MU k by the cloud. We consider two models of scaling for the computational capability of the cloud. In the *elastic* model each MU that offloads receives $F_k^c = F^c$ amount of computing power, which is a reasonable assumption for large cloud computing infrastructures. In the *non-elastic* model an MU that offloads is assigned $F_k^c(\mathbf{d}) = F^c/n(\mathbf{d})$ computation capability, i.e., the computing power is shared equally among all MUs that offload, which may be a reasonable model of emerging mobile edge cloud infrastructures with limited computational power and scaling [5].

Given F_k^c we use a linear model to compute the execution time of a task $\langle D_k, L_k \rangle$ that is offloaded by MU k ,

$$T_k^{c,exe} = \frac{L_k}{F_k^c}. \quad (6)$$

Figure 1 shows an example of a mobile cloud computing system that consists of $I = 3$ APs and $K = 5$ MUs in which MUs a and c offload using AP 1, MU b offloads using AP 2, and MUs d and e perform the local computation.

C. Cost Model

We consider that the cost of an MU can be modeled as a linear combination of the time it takes to finish the computation and its energy consumption. For MU k we denote by γ_k^E the weight attributed to energy consumption and by γ_k^T the weight attributed to the time it takes to finish the computation, $0 \leq \gamma_k^E < \gamma_k^T \leq 1$.

Using these notation, for the case of local computing the cost of MU k is determined by the local computing

time and the corresponding energy consumption,

$$C_k^0 = \gamma_k^T T_k^0 + \gamma_k^E E_k^0 = \left(\frac{\gamma_k^T}{F_k^0} + \gamma_k^E v_k \right) L_k. \quad (7)$$

For the case of offloading the cost is determined by the transmission time, the corresponding transmit energy, and the computing time in the cloud,

$$\begin{aligned} C_{k,i}^c(\mathbf{d}) &= \gamma_k^T (T_k^{c,exe} + T_{k,i}^{c,off}(\mathbf{d})) + \gamma_k^E E_{k,i}^c(\mathbf{d}) \\ &= (\gamma_k^T + \gamma_k^E P_k) \frac{D_k}{R_k^i(\mathbf{d})} + \gamma_k^T \frac{L_k}{F_k^c}. \end{aligned} \quad (8)$$

Similar to previous works [7], [13], [14], we do not model the time needed to transmit the results of the computation from the cloud server to the MU, as for typical applications like face and speech recognition, the size of the result of the computation is much smaller than D_k .

For notational convenience let us define the indicator function $I(d_k, i)$ for MU k as

$$I(d_k, i) = \begin{cases} 1, & \text{if } d_k = i \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

We can then express the cost of MU k in strategy profile \mathbf{d} as

$$C_k(\mathbf{d}) = C_k^0 I(d_k, 0) + \sum_{i \in \mathcal{I}} C_{k,i}^c(\mathbf{d}) I(d_k, i). \quad (10)$$

D. Computation Offloading Game

We consider that the objective of each MU is to minimize its cost (10), i.e., to find a strategy

$$d_k^* \in \arg \min_{d_k \in \mathcal{D}_k} C_k(d_k, d_{-k}), \quad (11)$$

where we use d_{-k} to denote the strategies of all MUs except MU k . Clearly, the strategy of an MU influences the cost of the other MUs, and thus we can model the problem as a strategic game $\Gamma = \langle \mathcal{K}, (\mathcal{D}_k)_k, (C_k)_k \rangle$, in which the players are the MUs. We refer to the game as the *computation offloading game*. We are interested in whether cost minimizing MUs can reach a strategy profile in which no MU can further decrease her cost through changing her strategy, i.e., a Nash equilibrium of the game Γ .

Definition 1. A Nash equilibrium (NE) of the strategic game $\langle \mathcal{K}, (\mathcal{D}_k)_k, (C_k)_k \rangle$ is a strategy profile d^* such that

$$C_k(d_k^*, d_{-k}^*) \leq C_k(d_k, d_{-k}^*).$$

Given a strategy profile (d_k, d_{-k}) we say that strategy d_k' is an improvement step for MU k if $C_k(d_k', d_{-k}) < C_k(d_k, d_{-k})$. We call a sequence of improvement steps in which one MU changes her strategy at a time an *improvement path*. Furthermore, we say that a strategy d_k^* is a best reply to d_{-k} if it solves (11), and we call an improvement path in which all improvement steps are best reply a *best improvement path*. Observe that in a NE all MUs play their best replies to each others' strategies.

In the rest of the paper we investigate whether NE exist for the *elastic* and for the *non-elastic* cloud model, and whether the MUs can compute a NE efficiently using distributed algorithms.

III. EQUILIBRIA IN CASE OF AN ELASTIC CLOUD

Recall that under the *elastic* cloud model the cloud computation capability assigned to user k is independent

of the other players' strategies, $F_k^c = F^c$. Thus, the cost function in the case of offloading can be expressed as

$$C_{k,i}^c(\mathbf{d}) = (\gamma_k^T + \gamma_k^E P_k) D_k \frac{n_i(\mathbf{d})}{B_i} + \gamma_k^T \frac{L_k}{F^c}. \quad (12)$$

We start with formulating an insightful structural result about the best responses of the MUs, which we will use later to prove the existence of NE.

Lemma 1. Given the strategy profile d_{-k} of the MUs other than k in the computation offloading game with elastic cloud, a best reply d_k^* of user k satisfies the following threshold strategy

$$d_k^* = \begin{cases} 0, & \text{if } M_k \leq \frac{n_i(i, d_{-k})}{B_i} \text{ for } \forall i \in \mathcal{I} \\ i, & \text{if } \frac{n_i(i, d_{-k})}{B_i} \leq \min \left\{ M_k, \min_{j \in \mathcal{I} \setminus \{i\}} \frac{n_j(j, d_{-k})}{B_j} \right\} \end{cases} \quad (13)$$

where

$$M_k = \frac{\gamma_k^E v_k + \gamma_k^T \left(\frac{1}{F_k^0} - \frac{1}{F^c} \right)}{\gamma_k^T + \gamma_k^E P_k} \cdot \frac{L_k}{D_k}.$$

Proof: Based on (7), (9), (10) and (12), the cost of MU k when choosing d_k is

$$\begin{aligned} C_k(d_k, d_{-k}) &= C_k^0 I(d_k, 0) + \sum_{i=1}^I C_{k,i}^c(d_k, d_{-k}) I(d_k, i) \\ &= \left(\frac{\gamma_k^T}{F_k^0} + \gamma_k^E v_k \right) L_k I(d_k, 0) \\ &\quad + \sum_{i=1}^I \left((\gamma_k^T + \gamma_k^E P_k) D_k \frac{n_i(i, d_{-k})}{B_i} + \gamma_k^T \frac{L_k}{F^c} \right) I(d_k, i). \end{aligned}$$

Let us first consider the case that the best reply of MU k is $d_k^* = 0$. We then have that $C_k(0, d_{-k}) \leq C_k(i, d_{-k})$ for every AP $i \in \mathcal{I}$, which implies that

$$\left(\frac{\gamma_k^T}{F_k^0} + \gamma_k^E v_k \right) L_k \leq (\gamma_k^T + \gamma_k^E P_k) D_k \frac{n_i(i, d_{-k})}{B_i} + \gamma_k^T \frac{L_k}{F^c}.$$

After algebraic manipulations we obtain

$$M_k \triangleq \frac{\gamma_k^E v_k + \gamma_k^T \left(\frac{1}{F_k^0} - \frac{1}{F^c} \right)}{\gamma_k^T + \gamma_k^E P_k} \cdot \frac{L_k}{D_k} \leq \frac{n_i(i, d_{-k})}{B_i}.$$

Let us now consider the case when the best reply of MU k is $d_k^* = i$. We then have that $C_k(i, d_{-k}) \leq C_k(0, d_{-k})$ and $C_k(i, d_{-k}) \leq C_k(j, d_{-k})$ for every AP $j \in \mathcal{I} \setminus \{i\}$. Following the same reasoning as above, it is easy to see that $C_k(i, d_{-k}) \leq C_k(0, d_{-k}^*)$ implies that $\frac{n_i(i, d_{-k})}{B_i} \leq M_k$. It is easy to see that $n_j(j, d_{-k}) = n_j(i, d_{-k}) + 1$, and thus $C_k(i, d_{-k}^*) \leq C_k(j, d_{-k}^*)$ implies that

$$(\gamma_k^T + \gamma_k^E P_k) D_k \frac{n_i(i, d_{-k})}{B_i} \leq (\gamma_k^T + \gamma_k^E P_k) D_k \frac{n_j(j, d_{-k})}{B_j}$$

which is equivalent to

$$\frac{n_i(i, d_{-k})}{B_i} \leq \frac{n_j(j, d_{-k})}{B_j}.$$

The above threshold strategy allows players to compute their best and better replies efficiently. In what follows we show that the computation offloading game with elastic cloud admits a NE, and a NE can be computed by iterative computation of the players' better or best replies, i.e.,

following an improvement path. Before we formulate the theorem, let us recall the definition of a generalized ordinal potential from [15].

Definition 2. A function $\Phi : \times \mathcal{D}_k \rightarrow \mathbb{R}$ is a generalized ordinal potential function for the strategic game $\langle \mathcal{K}, (\mathcal{D}_k)_k, (C_k)_k \rangle$ if for an arbitrary strategy profile (d_k, d_{-k}) and for any corresponding improvement step d'_k it holds that

$$C_k(d'_k, d_{-k}) - C_k(d_k, d_{-k}) < 0 \Rightarrow \Phi(d'_k, d_{-k}) - \Phi(d_k, d_{-k}) < 0.$$

Theorem 1. The computation offloading game with elastic cloud admits the generalized ordinal potential function

$$\Phi(\mathbf{d}) = \sum_{m=1}^I \sum_{n=1}^{n_m(\mathbf{d})} \frac{n}{B_m} + \sum_{s=1}^K M_s I(d_s, 0), \quad (14)$$

and hence it possesses a pure strategy Nash equilibrium.

Proof: To prove that $\Phi(\mathbf{d})$ is a generalized ordinal potential function, we first show that $C_k(i, d_{-k}) < C_k(0, d_{-k})$ implies $\Phi_k(i, d_{-k}) < \Phi_k(0, d_{-k})$ for a MU k . According to (7), (10) and (12), the condition $C_k(i, d_{-k}) < C_k(0, d_{-k})$ implies that

$$\frac{n_i(i, d_{-k})}{B_i} < M_k \quad (15)$$

for the strategy profile (i, d_{-k}) it holds that

$$\Phi(i, d_{-k}) = \sum_{n=1}^{n_i(i, d_{-k})} \frac{n}{B_i} + \sum_{m \neq i} \sum_{n=1}^{n_m(i, d_{-k})} \frac{n}{B_m} + \sum_{s \neq k} M_s I(d_s, 0),$$

and for the strategy profile $(0, d_{-k})$

$$\Phi(0, d_{-k}) = \sum_{n=1}^{n_i(0, d_{-k})} \frac{n}{B_i} + \sum_{m \neq i} \sum_{n=1}^{n_m(0, d_{-k})} \frac{n}{B_m} + M_k + \sum_{s \neq k} M_s I(d_s, 0).$$

Since $n_i(i, d_{-k}) = n_i(0, d_{-k}) + 1$, we obtain

$$\Phi(i, d_{-k}) - \Phi(0, d_{-k}) = \frac{n_i(i, d_{-k})}{B_i} - M_k.$$

It follows from (15) that $\Phi(i, d_{-k}) - \Phi(0, d_{-k}) < 0$. Similarly, we can show that $C_k(0, d_{-k}) < C_k(i, d_{-k})$ implies $\Phi_k(0, d_{-k}) < \Phi_k(i, d_{-k})$.

Second, we show that $C_k(i, d_k) < C_k(j, d_k)$ implies $\Phi_k(i, d_k) < \Phi_k(j, d_k)$ for a MU k . According to (10) and (12), the condition $C_k(i, d_k) < C_k(j, d_k)$ implies that

$$\frac{n_i(i, d_k)}{B_i} < \frac{n_j(j, d_k)}{B_j} \quad (16)$$

Let us rewrite Φ by separating the terms for APs i and j ,

$$\begin{aligned} \Phi(i, d_{-k}) &= \sum_{n=1}^{n_i(i, d_{-k})} \frac{n}{B_i} + \sum_{n=1}^{n_j(i, d_{-k})} \frac{n}{B_j} \\ &+ \sum_{m \neq i, j} \sum_{n=1}^{n_m} \frac{n}{B_m} + \sum_{s \neq k} M_s I(d_s, 0). \end{aligned} \quad (17)$$

Since $n_i(i, d_{-k}) = n_i(j, d_{-k}) + 1$ and $n_j(j, d_{-k}) = n_j(i, d_{-k}) + 1$, we have that

$$\Phi(i, d_{-k}) - \Phi(j, d_{-k}) = \frac{n_i(i, d_{-k})}{B_i} - \frac{n_j(j, d_{-k})}{B_j}.$$

It follows from (16) that $\Phi(i, d_{-k}) - \Phi(j, d_{-k}) < 0$, which

proves the theorem. \blacksquare

The existence of a generalized ordinal potential allows us to formulate a simple algorithm for computing a Nash equilibrium by leveraging the fact that in a game that all improvement paths are finite, i.e., lead to a Nash equilibrium, in a finite strategic game that admits a generalized ordinal potential function [15].

Corollary 1. Starting from an arbitrary initial strategy profile, let one MU at a time perform an improvement step iteratively. The algorithm terminates in a NE after a finite number of steps for the computation offloading game with elastic cloud.

IV. EQUILIBRIA IN CASE OF A NON-ELASTIC CLOUD

In the case of a non-elastic cloud the computation capability F_k^c that is assigned to MU k in the cloud server depends on the other MUs' strategies, and thus the cost function in case of offloading can be expressed as

$$C_{k,i}^c(\mathbf{d}) = (\gamma_k^T + \gamma_k^E P_k) D_k \frac{n_i(\mathbf{d})}{B_i} + \gamma_k^T \frac{L_k}{F^c} n(\mathbf{d}). \quad (18)$$

A natural question is whether a generalized ordinal potential similar to (14) exists in the case of non-elastic cloud, in which case all improvement paths would be finite. We first show that if we only allow MUs to change between APs, but we do not allow them to start or to stop offloading, then this holds.

Lemma 2. Consider an arbitrary strategy profile \mathbf{d} , and consider that (i) the improvement step d'_k of MU $k \in O(\mathbf{d})$ is constrained to $d'_k \in \mathcal{I}$, and (ii) MUs $k \notin O(\mathbf{d})$ are not allowed to perform improvement steps. Then all improvement paths that satisfy constraints (i) and (ii) are finite.

Proof: First, observe that the set $O(\mathbf{d})$ of offloaders is unchanged during an improvement path with constraints (i) and (ii). For a strategy profile \mathbf{d} let vector $\gamma(\mathbf{d}) \in \mathbb{R}_{\geq 0}^{|O(\mathbf{d})|}$ contain the cost $C_k(\mathbf{d})$ for MUs $k \in O(\mathbf{d})$ in decreasing order. Let $\mathbf{d}' = (d'_k, d_{-k})$ be the strategy profile after an improvement step made by MU k that satisfies constraint (i), and let $i = d_k$ and $j = d'_k$. Since $\frac{n_j(\mathbf{d})+1}{B_j} < \frac{n_i(\mathbf{d})}{B_i}$ must hold for the change of APs to be an improvement step, we have $\gamma(\mathbf{d}') \prec_L \gamma(\mathbf{d})$, where \prec_L stands for lexicographically smaller. Since $\gamma(\mathbf{d})$ decreases in the lexicographical sense upon every improvement step, and the number of strategy profiles is finite, the improvement paths must be finite. \blacksquare

Thus, if MUs can only change between APs, they terminate after a finite number of improvement steps. Unfortunately, as the following example shows, this is not the case if MUs can decide not to offload, and thus the algorithm in Corollary 1 cannot be used to compute a NE, even if a NE exists.

Example 1. Consider a computation offloading game with non-elastic cloud where $\mathcal{K} = \{a, b, c, d, e\}$ and $\mathcal{I} = \{1, 2, 3\}$ as illustrated in Figure 1. Figure 2 shows a cyclic improvement path starting from the strategy profile $(1, 2, 1, 0, 0)$, in which MUs a and c are connected to AP 1, MU b is connected to AP 2 and MUs d and e perform local computation.

d_k	d_a	d_b	d_c	d_d	d_e
$\mathbf{d}(0)$	1	2	1	0	0
$\mathbf{d}(1)$	1	2	↓	0	0
$\mathbf{d}(2)$	1	↓	2	0	0
$\mathbf{d}(3)$	1	0	2	↓	0
$\mathbf{d}(4)$	1	0	2	2	↓
$\mathbf{d}(5)$	1	0	↓	2	2
$\mathbf{d}(6)$	1	↓	3	2	2
$\mathbf{d}(7)$	1	3	1	2	↓
$\mathbf{d}(8)$	1	3	1	↓	0
$\mathbf{d}(9)$	1	↓	2	1	0

Fig. 2. A cyclic improvement path in a computation offloading game with non-elastic cloud, 3 APs and 5 MUs. Rows correspond to strategy profiles, columns to MUs. An arrow between adjacent rows indicates the MU that performs the improvement step. The cycle consists of 9 improvement steps, and involves some MUs to start and to stop offloading. The inequalities on the right show the condition under which the change of strategy is an improvement step.

Starting from the initial strategy profile $(1, 2, 1, 0, 0)$, Player c revises its strategy to AP 2, which is an improvement step if $B_2 > B_1$, as shown in inequality (1) in the figure. Observe that after 9 improvement steps the players reach the initial strategy profile. For each step the inequality on the right provides the condition for being an improvement. It follows from inequalities (1), (5) and (9) that $B_2 > B_1$, $B_1 > \frac{2}{3}B_2$ and $B_2 > B_3$, respectively. Since, $\frac{1}{B_3}(\gamma_b^T + \gamma_b^E P_b)D_b + 5\gamma_b^T \frac{L_b}{F^c} > \frac{1}{B_3}(\gamma_b^T + \gamma_b^E P_b)D_b + 3\gamma_b^T \frac{L_b}{F^c}$ holds, from inequalities (2) and (6) follows that $B_3 > \frac{1}{2}B_2$. Combining inequalities (3) and (8) we have that $\gamma_d^T \frac{L_d}{F^c} > \frac{1}{B_2}(\gamma_d^T + \gamma_d^E P_d)D_d$. Similarly, it follows from inequalities (4) and (7) that $\gamma_e^T \frac{L_e}{F^c} > \frac{1}{B_2}(\gamma_e^T + \gamma_e^E P_e)D_e$. Given these constraints, an instance of the example can be formulated easily.

An important consequence of the cycle in the improvement path is that the computation offloading game with non-elastic cloud does not allow a potential function, and thus Corollary 1 cannot be applied. Yet, as we now show, NE always exist.

Theorem 2. *The computation offloading game with non-elastic cloud possesses a pure strategy Nash equilibrium.*

Proof: We use induction in the number K of players in order to prove the theorem, and we denote by $K^{(t)} = t$ the number of MUs that are involved in the game in induction step t .

It is clear that for $K^{(1)} = 1$ there is a NE, in which the only participating MU plays her best reply $d_k^*(1)$. Since there are no other MUs, $\mathbf{d}^*(1)$ is a NE. Observe that if $d_k^*(1) = 0$, MU k would never have an incentive to deviate from this decision, because the number of players that offload will not decrease as more MUs are added. Otherwise, if MU k decides to offload, her best reply is

given by $d_k^*(1) = \arg \max_{i \in \mathcal{I}} B_i$.

Assume now that for $t - 1 > 0$ there is a NE $\mathbf{d}^*(t - 1)$. Upon induction step t one MU enters the game; we refer to this MU as MU $K^{(t)}$. Let MU $K^{(t)}$ play her best reply $d_{K^{(t)}}^*(t)$ with respect to the NE strategy profile of the MUs that already participated in induction step $t - 1$, i.e., with respect to $d_{-K^{(t)}}(t) = \mathbf{d}^*(t - 1)$. After that, MUs can perform best improvement steps one at a time starting from the strategy profile $\mathbf{d}(t) = (d_{K^{(t)}}^*(t), d_{-K^{(t)}}(t))$, following the algorithm shown in Figure 3. We refer to this as the update phase. In order to prove that there is a NE in induction step t , in the following we show that the MUs will perform a finite number of best improvement steps in the update phase.

Let us define the reluctance to offload via AP i of MU k in a strategy profile $\mathbf{d}(t)$ as $\rho_k(\mathbf{d}(t)) = \frac{C_{k,i}(\mathbf{d}(t))}{C_k^0}$, and let us rank the MUs that play the same strategy in decreasing order of reluctance. We use the triplet (t, l, i) to index the MU that in step t occupies position l in the ranking for AP i , i.e., $\rho_{(t,1,i)}(\mathbf{d}(t)) \geq \rho_{(t,2,i)}(\mathbf{d}(t)) \geq \dots \geq \rho_{(t,n_i(\mathbf{d}(t)),i)}(\mathbf{d}(t))$. Note that for AP i it is MU $(t, 1, i)$ that can gain most by changing her strategy among all MUs $k \in O_i(\mathbf{d}(t))$.

Observe that if $d_{K^{(t)}}^*(t) = 0$, then $n_i(\mathbf{d}(t)) = n_i(\mathbf{d}^*(t - 1))$ for every $i \in \mathcal{I}$ and thus $\mathbf{d}(t)$ is a NE. If $d_{K^{(t)}}^*(t) = i \in \mathcal{I}$, but none of the MUs want to deviate from their strategy in $\mathbf{d}^*(t - 1)$ then $\mathbf{d}(t)$ is a NE. Otherwise, we can have one or both of the following: (i) for some MUs $k \in O_i(\mathbf{d}(t))$ offloading using AP i is not a best reply anymore, (ii) for some MUs $k \in O_j(\mathbf{d}(t))$ for $j \in \mathcal{I} \setminus \{i\}$, offloading using AP j is not a best reply anymore. Let us denote by $D_{O \rightarrow L}$ the set of APs with at least one MU that wants to deviate from her strategy either for (i) or for (ii).

Observe that case (i) can happen only if $\rho_{(t,1,i)}(\mathbf{d}(t)) > \rho_{K^{(t)}}(\mathbf{d}(t))$, as otherwise no MU $k \in O_i(\mathbf{d}(t))$ would be able to gain by changing her strategy from AP i . Now, since $d_{K^{(t)}}(t) = i$ it follows that $\frac{n_i(\mathbf{d}^*(t-1))+1}{B_i} \leq \frac{n_j(\mathbf{d}^*(t-1))+1}{B_j}$ for every $j \in \mathcal{I} \setminus \{i\}$. Therefore, in case (i) an MU $k \in O_i(\mathbf{d}(t))$ cannot decrease her offloading cost by choosing another AP j ; as an improvement step she would change her strategy to local computing. Let now MU $(t, 1, i)$ perform an improvement step, and let us denote the resulting strategy profile by $\mathbf{d}'(t)$ (Line 4). Since MU $(t, 1, i)$ changed from AP i to local computation, $n_i(\mathbf{d}'(t)) = n_i(\mathbf{d}^*(t - 1))$ would hold for every $i \in \mathcal{I}$ and $\mathbf{d}'(t)$ would be a NE.

Let us now consider case (ii). The only reason why case (ii) could happen is that the number of players that offload was incremented, i.e., $n(\mathbf{d}(t)) = n(\mathbf{d}^*(t - 1)) + 1$. Thus, the best improvement of every MU $k \in O_j(\mathbf{d}(t))$ that wants to deviate would be to perform the computation locally. Among all MUs that would like to deviate, let us choose the MU with highest reluctance $\rho_k(\mathbf{d}(t))$ (note that this is MU $(t, 1, j)$ for some $j \neq i$), and let her perform the improvement step, i.e., change to local computation (Lines 9 – 12). Let the resulting strategy profile be $\mathbf{d}'(t)$. Due to this improvement step $n_j(\mathbf{d}'(t)) = n_j(\mathbf{d}^*(t - 1)) - 1$, and thus some MUs may be able to decrease their cost by connecting to AP j . If there is no MU $k \in \mathcal{K} \setminus O(\mathbf{d}'(t))$ that would like to start offloading, then there is no more

```

1: if  $i \in D_{O \rightarrow L}$  then
2:   /* Corresponds to case (i) */
3:   Let  $k' \leftarrow (t, 1, i)$ 
4:   Let  $\mathbf{d}'(t) = (0, d_{-k'}(t))$  /* Best reply by MU  $k'$  */
5: else if  $j \in D_{O \rightarrow L}$  then
6:   /* Corresponds to case (ii) */
7:   Let  $\mathbf{d}'(t) = \mathbf{d}(t)$ 
8:   while  $D_{O \rightarrow L} \neq \emptyset$  do
9:      $j \leftarrow \arg \max_{j' \in D_{O \rightarrow L}} \rho_{(t, 1, j')}(\mathbf{d}'(t))$ 
10:    /* AP with MU with highest reluctance */
11:    Let  $k' \leftarrow (t, 1, j)$ 
12:    Let  $\mathbf{d}'(t) = (0, d'_{-k'}(t))$ 
13:    /* Best reply by MU  $(t, 1, j)$  */
14:     $D_{L \rightarrow O} = \{k | d'_k(t) = 0, C_k^0 \geq C_{k,j}^c(\mathbf{d}'(t))\}$ 
15:    if  $D_{L \rightarrow O} \neq \emptyset$  then
16:       $k' \leftarrow \arg \max_{k \in D_{L \rightarrow O}} C_k^0$ 
17:      /* MU with highest local cost */
18:      Let  $\mathbf{d}'(t) = (j, d'_{-k'}(t))$ 
19:      /* Best reply by MU  $k'$  */
20:       $D_{O \rightarrow L} = \{j \in \mathcal{I} | \exists k \in O_j(\mathbf{d}'(t)), C_{k,j}^c(\mathbf{d}'(t)) \geq C_k^0\}$ 
21:    else
22:       $D_{O \rightarrow O} = \{i | i \in \mathcal{I} \setminus \{j\}, \frac{n_j(\mathbf{d}'(t))+1}{B_j} < \frac{n_i(\mathbf{d}'(t))}{B_i}\}$ 
23:      while  $D_{O \rightarrow O} \neq \emptyset$  do
24:         $i \leftarrow \arg \max_{i' \in D_{O \rightarrow O}} \rho_{(t, 1, i')}(\mathbf{d}'(t))$ 
25:        /* AP with MU with highest reluctance */
26:        Let  $k' \leftarrow (t, 1, i)$ 
27:        Let  $\mathbf{d}'(t) = (j, d'_{-k'}(t))$ 
28:        /* Best reply by MU  $(t, 1, i)$  */
29:        Let  $j \leftarrow i$ 
30:         $D_{O \rightarrow O} = \{i | i \in \mathcal{I} \setminus \{j\}, \frac{n_j(\mathbf{d}'(t))+1}{B_j} < \frac{n_i(\mathbf{d}'(t))}{B_i}\}$ 
31:      end while
32:    end if
33:  end while
34: end if

```

Fig. 3. Pseudo code of the update phase of the distributed algorithm.

MU that would like to stop offloading either because $n(\mathbf{d}'(t)) = n(\mathbf{d}^*(t-1)) - 1$. Otherwise, among all MUs $k \in \mathcal{K} \setminus O(\mathbf{d}'(t))$ that would like to start offloading, let MU k' with highest local computing cost $C_{k'}^0$ perform an improvement step, i.e., connect to AP j . We now repeat these steps starting from Line 8 until no more MU wants to stop offloading. This iteration will stop after a finite number of steps, as the MU that stops offloading always has higher reluctance than that one that replaces it, and the number of MUs is finite. Let j be the AP that the last MU that stopped offloading was connected to. If the last MU that stopped offloading was replaced by an MU that did not offload before, then we reached a NE. Otherwise some MUs may want to change to AP j . By Lemma 2 if we only allow MUs to change between APs, we terminate in a finite number of improvement steps. Now, no MU wants to stop offloading, and no MU wants to start offloading either, because they did not want to do so before the MUs were allowed to change APs. Hence we reached a NE, which proves the inductive step. ■

As we next show, the above constructive proof provides a low complexity algorithm for computing a Nash equilibrium of the game.

Proposition 3. *For the computation offloading game with non-elastic cloud, when player $K^{(t)}$ enters the game in equilibrium $\mathbf{d}^*(t-1)$, a new Nash equilibrium can be computed in $O(K^{(t)} + I)$ time.*

Proof: Let us consider inductive step t in which MU $K^{(t)}$ enters the game. From the proof of Theorem 2 it follows that if $d_{K^{(t)}}^*(t) = 0$, or if $d_{K^{(t)}}^*(t) = i \in \mathcal{I}$ but none of the MUs want to deviate from their strategy in $\mathbf{d}^*(t-1)$, then a NE is reached without any update steps. If $d_{K^{(t)}}^*(t) = i \in \mathcal{I}$ and case (i) happens, a NE is reached after one update step. Now let us consider that $d_{K^{(t)}}^*(t) = i \in \mathcal{I}$ and case (ii) happens. Note that from Theorem 2 it follows that case (ii) can happen only if $I > 1$. In what follows we characterize the longest sequences of update steps that lead to a NE for the case when $K^{(t)}$ is even and when it is odd.

If $K^{(t)}$ is even, the worst case scenario is when $|O(\mathbf{d}^*(t-1))| = \lceil \frac{K^{(t)}-1}{2} \rceil$ and $n_i(\mathbf{d}^*(t-1)) = 0$, in the new strategy profile $\mathbf{d}(t)$ every MU $k \in O(\mathbf{d}^*(t-1))$ wants to change to local computing, and when MU k with highest reluctance $\rho_k(\mathbf{d}(t))$ changes to local computing, all MUs $k \in \mathcal{K} \setminus O(\mathbf{d}^*(t-1))$, i.e., a total of $\lfloor \frac{K^{(t)}-1}{2} \rfloor$ MUs would like to start offloading. In the corresponding sequence of update steps that leads to a NE, in the first $2\lfloor \frac{K^{(t)}-1}{2} \rfloor + 1$ update steps all MUs $k \in O(\mathbf{d}^*(t-1))$ stop to offload and all MUs $k \in \mathcal{K} \setminus O(\mathbf{d}^*(t-1))$ start to offload, and in the next $(I-1)$ update steps $(I-1)$ MUs change between APs. Therefore, a NE is reached after at most $2\lfloor \frac{K^{(t)}-1}{2} \rfloor + 1 + (I-1)$ update steps.

If $K^{(t)}$ is odd, the worst case scenario is when $n_i(\mathbf{d}^*(t-1)) = 1$, in the new strategy profile $\mathbf{d}(t)$ a total of $\lfloor \frac{K^{(t)}-1}{2} \rfloor$ MUs of the $|O(\mathbf{d}^*(t-1))| = \lfloor \frac{K^{(t)}-1}{2} \rfloor + 1$ MUs that offload want to change to local computing, and when MU k with highest reluctance $\rho_k(\mathbf{d}(t))$ changes to local computing, all MUs $k \in \mathcal{K} \setminus O(\mathbf{d}^*(t-1))$, i.e., a total of $\lfloor \frac{K^{(t)}-1}{2} \rfloor - 1$ MUs would like to start offloading. Following the same reasoning as above, we obtain that a NE is reached after at most $2(\lfloor \frac{K^{(t)}-1}{2} \rfloor - 1) + 1 + (I-1)$ update steps. ■

Consider now that we add players one at a time, we then obtain the following bound on the complexity of computing a NE.

Corollary 2. *A Nash equilibrium of the computation offloading game with non-elastic cloud can be computed in $O(K^2 + KI)$ time.*

So far we have shown that starting from a NE and adding a new player, a new NE can be computed. We now show a similar result for the case when a player leaves.

Theorem 4. *Consider the computation offloading game with non-elastic cloud, and assume the system is in a NE. If an existing player leaves the game and the remaining players play best replies, they converge to a Nash equilibrium after a finite number of updates.* ■

Proof: Let us consider that player k leaves the game, when the system is in a NE. If the strategy of player k is to perform local computation, none of the remaining players would be affected when player k leaves. If the strategy of player k is to offload using one of the APs, we can consider player k as a player that after changing his strategy from offloading to local computing, would have no incentive to offload again. Recall from the proof of Theorem 2 that when a player changes her strategy from offloading to local computing the game converges to a Nash equilibrium after a finite number of updates. This proves the theorem. ■

Observe that Theorem 2 and Theorem 4 allow for efficient computation of equilibrium system operation if the time between user arrivals and departures is sufficient to compute a new equilibrium. Furthermore, the computation can be done in a decentralized manner, by letting MUs perform best improvements one at a time. The advantage of such a decentralized implementation could be that MUs do not have to reveal their parameters.

V. PRICE OF ANARCHY

We have so far shown that NE exist and provided low complexity algorithms for computing a NE. We now address the important question how far the system performance would be from optimal in a NE. To quantify the difference from the optimal performance we use the price of anarchy (PoA), defined as the ratio of the worst case NE cost and the minimal cost

$$PoA = \frac{\max_{\mathbf{d}^*} \sum_{k \in \mathcal{K}} C_k(\mathbf{d}^*)}{\min_{\mathbf{d} \in \mathcal{D}} \sum_{k \in \mathcal{K}} C_k(\mathbf{d})}. \quad (19)$$

In what follows we give an upper bound on the PoA.

Theorem 5. *The price of anarchy for the computation offloading game is upper bounded by*

$$\frac{\sum_{k \in \mathcal{K}} C_k^0}{\sum_{k \in \mathcal{K}} \min\{C_k^0, C_{k,1}^c, \dots, C_{k,I}^c\}},$$

both in the case of elastic cloud and in the case of non-elastic cloud.

Proof: First we show that if there is a NE in which all players perform local computation then it is the worst case NE. To show this let \mathbf{d}^* be an arbitrary NE. Observe that $C_k(d_k^*, d_{-k}^*) \leq C_k^0$ holds for every player $k \in \mathcal{K}$. Otherwise, if $\exists k \in \mathcal{K}$ such that $C_k(d_k^*, d_{-k}^*) > C_k^0$, player k would have an incentive to deviate from decision d_k^* , which contradicts our initial assumption that \mathbf{d}^* is a NE. Thus in any NE $\sum_{k \in \mathcal{K}} C_k(d_k^*, d_{-k}^*) \leq \sum_{k \in \mathcal{K}} C_k^0$ holds, and if all players performing local computation is a NE then it is the worst case NE.

Now we derive a lower bound for the optimal solution of the computation offloading game in the case of both the elastic and non-elastic cloud. Let us consider an arbitrary decision profile $(d_k, d_{-k}) \in \mathcal{D}$. If $d_k = 0$, then $C_k(d_k, d_{-k}) = C_k^0$. Otherwise, if $d_k = i$ for some $i \in \mathcal{I}$, we have that in the best case $d_{k'} = 0$ for every $k' \in \mathcal{K} \setminus \{k\}$, and thus $n(\mathbf{d}) = 1$. Therefore,

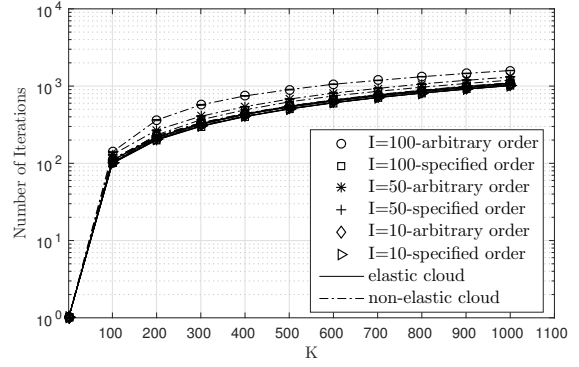


Fig. 6. Number of iterations vs. number of users K for the elastic and non-elastic cloud, $I = 10, 50$ and 100 . The results shown are the averages of 100 simulations, together with 95% confidence intervals.

$R_k^i(d_k, d_{-k}) \leq B_i$ and $F_k^c \leq F^c$, which implies that

$$\begin{aligned} C_{k,i}^c(d_k, d_{-k}) &= (\gamma_k^T + \gamma_k^E P_k) \frac{D_k}{R_k^i(d_k, d_{-k})} + \gamma_k^T \frac{L_k}{F_k^c} \\ &\geq (\gamma_k^T + \gamma_k^E P_k) \frac{D_k}{B_i} + \gamma_k^T \frac{L_k}{F^c} = C_{k,i}^{\bar{c}}. \end{aligned}$$

Hence, we have $C_k(d_k, d_{-k}) \geq \min\{C_k^0, C_{k,1}^{\bar{c}}, \dots, C_{k,I}^{\bar{c}}\}$ and $\sum_{k \in \mathcal{K}} C_k(d_k, d_{-k}) \geq \sum_{k \in \mathcal{K}} \min\{C_k^0, C_{k,1}^{\bar{c}}, \dots, C_{k,I}^{\bar{c}}\}$. Using these we can establish the following bound

$$PoA = \frac{\max_{\mathbf{d}^*} \sum_{k \in \mathcal{K}} C_k(\mathbf{d}^*)}{\min_{\mathbf{d} \in \mathcal{D}} \sum_{k \in \mathcal{K}} C_k(\mathbf{d})} \leq \frac{\sum_{k \in \mathcal{K}} C_k^0}{\sum_{k \in \mathcal{K}} \min\{C_k^0, C_{k,1}^{\bar{c}}, \dots, C_{k,I}^{\bar{c}}\}},$$

which proves the theorem. ■

VI. NUMERICAL RESULTS

We use simulations to evaluate the cost performance and the computational time of the proposed distributed algorithms.

A. Evaluation Scenario

In all configurations, we consider that the bandwidth of each AP is drawn from a normal distribution with mean $\mu = 5$ MHz and standard deviation of 0.2μ . The parameters $\langle D_k, L_k \rangle$ that characterize the computation tasks, the computational capability of MU F_k^0 and the weights attributed to energy consumption γ_k^E and the time it takes to finish the computation γ_k^T were drawn from a continuous uniform distribution with parameters $[0.42, 2]$ Mb, $[0.1, 0.8]$ Gigacycles, $[0.5, 1]$ Gigacycles, $[0, 1]$ and $[0, 1]$, respectively. The consumed energy per CPU cycle v_k was set to $10^{-11} (F_k^0)^2$ according to measurements reported in [4], [16]. The data transmit power P_k was set to $0.4W$ according to [17]. In the case of the non-elastic cloud, the computation capability of the cloud F^c was set to 100 Gigacycles [18] and in the case of the elastic cloud each MU that offloads receives $F_k^c = 100$ Gigacycles amount of computing power.

In order to evaluate the cost performance of the equilibrium strategy profile \mathbf{d}^* computed by the proposed distributed algorithms, we computed the optimal strategy profile $\bar{\mathbf{d}}$ that minimizes the total cost, i.e., $\bar{\mathbf{d}} = \arg \min_{\mathbf{d}} \sum_{k \in \mathcal{K}} C_k(\mathbf{d})$. Furthermore, as a baseline for comparison we use the system cost that can be achieved in the strategy profile in which all MUs execute their computation tasks locally, which coincides with the bound on the PoA.

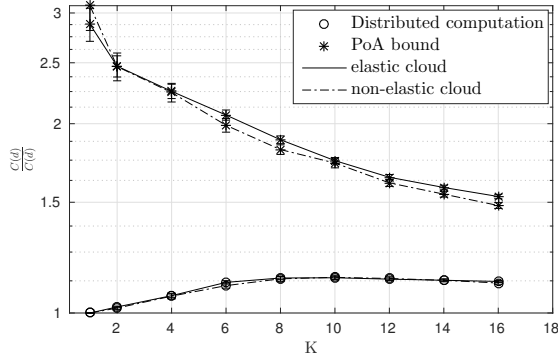


Fig. 4. Ratio between the total cost achieved by the proposed distributed algorithm and the optimal total cost for the elastic and non-elastic cloud, $I = 3$. The results shown are the averages of 500 simulations, together with 95% confidence intervals.

B. Price of Anarchy

Figure 4 shows the cost ratio $C(\mathbf{d}^*)/C(\bar{\mathbf{d}})$ in the case of the elastic as well as in the case of the non-elastic cloud. To make the computation of the optimal strategy profile $\bar{\mathbf{d}}$ feasible, we considered a scenario with $I = 3$ APs and we show the cost ratio $C(\mathbf{d}^*)/C(\bar{\mathbf{d}})$ as a function of the number of MUs. Figure 4 shows that the results reached by the algorithms are close to the optimal results and the difference between the elastic cloud case and the non-elastic cloud case is negligible. Furthermore, we can observe that the cost ratio increases slightly up to $K = 6$ MUs and from that point it remains fairly unchanged. This is due to the number of MUs that choose to offload, as we will see later. The upper bound on the PoA, which is also shown in Figure 4, additionally confirms that the proposed distributed algorithms perform good in terms of the cost ratio. It is interesting to note that the gap between the PoA bound and the actual cost ratio decreases with increasing number of MUs.

To get insight into the structure of the equilibrium strategy profile \mathbf{d}^* computed by the distributed algorithms, we compare the number of MUs that offload in equilibrium and the number of MUs that offload in the optimal strategy profile $\bar{\mathbf{d}}$, by computing the offloading difference ratio $(n(\mathbf{d}^*) - n(\bar{\mathbf{d}}))/K$. Figure 5 shows the offloading difference ratio corresponding to the results shown in Figure 4. The results show that the offloading difference ratio is fairly small in the case of the elastic cloud as well as in the case of the non-elastic cloud. As the number of MUs increases, the offloading difference ratio increases too, which explains the increased cost ratio observed in Figure 4, as more offloaders reduce the achievable rate, which in turn leads to increased costs. The observation that the number of MUs that offload is higher in equilibrium than in the optimal solution is consistent with the theory of the tragedy of the commons in economic theory [19].

C. Computational Complexity

In order to evaluate the computational complexity of the proposed algorithms, we study the number of iterations needed to compute the strategy profile \mathbf{d}^* for three scenarios with $I = 10, 50, 100$ APs, respectively. For the elastic cloud the number of iterations is the number of update steps, while for the non-elastic cloud the number of iterations is the sum of the update steps over all induction

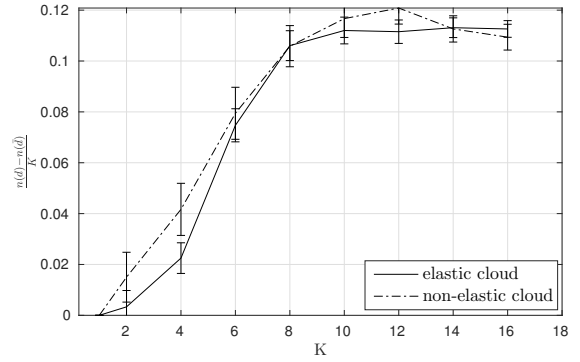


Fig. 5. Offloading ratio vs. number of users K for the elastic and non-elastic cloud, $I = 3$. The results shown are the averages of 500 simulations, together with 95% confidence intervals.

steps. Figure 6 shows the number of iterations as a function of the number of MUs. For the non-elastic cloud we consider two orderings of adding MUs: in the first case the MUs are added in random order, while in the second case the MUs enter the game in increasing order of their ratio $\frac{D_k}{C_k^0 L_k}$. In both cases we use the same simulation scenarios in order to compute the number of the iterations. Intuitively, we can expect that the second case results in a smaller number of the iterations, since the MUs with lower $\frac{D_k}{C_k^0 L_k}$ ratio have higher computational capability to execute computationally more demanding tasks with smaller offloading data size than the MUs with higher $\frac{D_k}{C_k^0 L_k}$ ratio. However, the simulation results show that the number of iterations is fairly insensitive to the order of adding the MUs and mostly depends on the number of MUs K . This insensitivity allows for the implementation of a very low-overhead decentralized solution, as the coordinator need not care about the order in which the MUs are added for computing the equilibrium allocation.

VII. RELATED WORK

There is a significant body of works that deals with the design of energy efficient computation offloading for a single mobile user [3], [4], [6], [20], [21], [13], [22]. The experimental results in [21] showed that significant battery power savings can be achieved by computation offloading. [6] studied the communication overhead of computation offloading and the impact of bandwidth availability on an experimental platform. [3] proposed a code partitioning solution for fine-grained energy-aware computation offloading. [13] proposed an algorithm for offloading partitioned code under bandwidth and delay constraints. [4] proposed CPU frequency and transmission power adaptation for energy-optimal computation offloading under delay constraints. [22] modeled the offloading problem under stochastic task arrivals as a Markov decision process and provided a near-optimal offloading policy.

A number of recent works considered the problem of joint energy minimization for multiple mobile users [8], [23], [9]. [8] studies computation partitioning for streaming data processing with the aim of maximizing throughput, considering sharing of computation instances among multiple mobile users, and proposes a genetic algorithm as a heuristic for solving the resulting optimization problem. [23] models computation offloading to a tiered

cloud infrastructure under user mobility in a location-time workflow framework, and proposes a heuristic for minimizing the users' cost. [9] aims at minimizing the mobile users' energy consumption by joint allocation of radio resources and cloud computing power, and provides an iterative algorithm to find a local minimum of the optimization problem.

A few recent works provided a game theoretic treatment of computation offloading in a game theoretical setting [24], [25], [7], [26], [27], [28]. [24] considers a two-stage problem, where first each mobile user decides what share of its task to offload so as to minimize its energy consumption and to meet its delay deadline, and then the cloud allocates computational resources to the offloaded tasks. [25] considers a two-tier cloud infrastructure and stochastic task arrivals and proves the existence of equilibria and provides an algorithm for computing and equilibrium. [27] considers tasks that arrive simultaneously, a single wireless link, and elastic cloud, and show the existence of equilibria when all mobile users have the same delay budget. Our work differs from [24] in that we consider that the allocation of cloud resources is known to the mobile users, from [25] in that we take into account contention in the wireless access, and from [27] in that we consider multiple wireless links and a non-elastic cloud.

Most related to our work are the problems considered in [7], [26], [28]. [7] considers contention on a single wireless link and an elastic cloud, assumes upload rates to be determined by the Shannon capacity of an interference channel, and shows that the game is a potential game. [26] extends the model to multiple wireless links and shows that the game is still a potential game. Unlike these works, we consider fair bandwidth sharing and consider the case of non-elastic cloud. [28] considers multiple wireless links, fair bandwidth sharing and a non-elastic cloud, and claims the game to have an exact potential. In our work we on the one hand extend the model to an elastic cloud, on the other hand we show that an exact potential cannot exist in case of a non-elastic cloud, but at the same time we prove the existence of an equilibrium allocation, provide an efficient algorithm with quadratic complexity for computing one, and provide a bound on the price of anarchy.

Besides providing efficient algorithms for computing equilibria, the importance of our contribution lies in the fact that while games with an elastic cloud are player-specific singleton congestion games for which the existence of equilibria is known [29], the non-elastic cloud model does not fall in this category of games and thus no general equilibrium existence result exists.

VIII. CONCLUSION

We have considered the problem of computation offloading in a multi-access wireless network by self-interested mobile users for mobile cloud computing, for the case of elastic and non-elastic cloud resources. We provided a game theoretical formulation of the problem, and showed that in the case of an elastic cloud a simple algorithm, in which users iteratively improve their allocations, can be used for computing an equilibrium. We showed that the same algorithm may fail in the case of a non-elastic cloud, but also showed that an equilibrium

always exists, and provided an algorithm for computing an equilibrium with quadratic complexity. Finally, we provided a bound on the price of anarchy. Simulation results show that the complexity bound is not tight, and the proposed algorithm scales better than quadratic in terms of the number of users, and the obtained equilibria provide good system performance.

REFERENCES

- [1] M. Hakkarainen, C. Woodward, and M. Billinghurst, "Augmented assembly using a mobile phone," in *Proc. of IEEE/ACM Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, Sept 2008, pp. 167–168.
- [2] J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uwave: Accelerometer-based personalized gesture recognition and its applications," in *Proc. of IEEE PerCom*, March 2009, pp. 1–9.
- [3] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "Maui: Making smartphones last longer with code offload," in *Proc. of ACM MobiSys*, 2010, pp. 49–62.
- [4] Y. Wen, W. Zhang, and H. Luo, "Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones," in *Proc. of IEEE INFOCOM*, March 2012, pp. 2716–2720.
- [5] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing: A key technology towards 5G," Sep. 2015.
- [6] M. V. Barbera, S. Kosta, A. Mei, and J. Stefa, "To offload or not to offload? The bandwidth and energy costs of mobile cloud computing," in *Proc. of IEEE INFOCOM*, April 2013, pp. 1285–1293.
- [7] X. Chen, "Decentralized computation offloading game for mobile cloud computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 4, pp. 974–983, 2015.
- [8] L. Yang, J. Cao, Y. Yuan, T. Li, A. Han, and A. Chan, "A framework for partitioning and execution of data stream applications in mobile cloud computing," *SIGMETRICS Perform. Eval. Rev.*, vol. 40, no. 4, pp. 23–32, Apr. 2013.
- [9] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 1, no. 2, pp. 89–103, Jun. 2015.
- [10] G. Iosifidis, L. Gao, J. Huang, and L. Tassiulas, "An iterative double auction for mobile data offloading," in *Proc. of WiOpt*, May 2013, pp. 154–161.
- [11] I. C. Wong, O. Oteri, and W. McCoy, "Optimal resource allocation in uplink sc-fdma systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 5, pp. 2161–2165, May 2009.
- [12] S. Cicalo and V. Tralli, "Fair resource allocation with qos support for the uplink of lte systems," in *Proc. of European Conference on Networks and Communications (EuCNC)*, Jun. 2015, pp. 180–184.
- [13] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 1991–1995, Jun. 2012.
- [14] K. Kumar and Y. H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *IEEE Computer Mag.*, vol. 43, no. 4, pp. 51–56, Apr. 2010.
- [15] D. Monderer and L. S. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, no. 1, pp. 124 – 143, 1996.
- [16] A. P. Miettinen and J. K. Nurminen, "Energy efficiency of mobile clients in cloud computing," in *Proc. of the 2nd USENIX Conf. Hot Topics Cloud Comput.*, 2010, pp. 4–4.
- [17] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy consumption in mobile phones: a measurement study and implications for network applications," in *Proc. of the 9th ACM SIGCOMM conference on Internet measurement conference*, ACM, 2009, pp. 280–293.
- [18] T. Soyata, R. Muralidharan, C. Funai, M. Kwon, and W. Heinzelman, "Cloud-vision: Real-time face recognition using a mobile-cloudlet-cloud acceleration architecture," in *IEEE Symp. on Computers and Communications (ISCC)*, 2012, pp. 59–66.
- [19] G. Hardin, "The tragedy of the commons," *Science*.
- [20] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mob. Netw. Appl.*, vol. 18, no. 1, pp. 129–140, Feb 2013.
- [21] A. Rudenko, P. Reiher, G. J. Popek, and G. H. Kuenning, "Saving portable computer battery power through remote process execution," *ACM Mob. Comput. Commun. Rev.*, vol. 2, no. 1, pp. 19–26, Jan 1998.

- [22] E. Hyttiä, T. Spyropoulos, and J. Ott, "Offload (only) the right jobs: Robust offloading using the Markov decision processes," in *Proc. of IEEE WoWMoM*, Jun. 2015, pp. 1–9.
- [23] M. R. Rahimi, N. Venkatasubramanian, and A. V. Vasilakos, "MuSIC: Mobility-aware optimal service allocation in mobile cloud computing," in *Proc. of IEEE CLOUD*, Jun. 2013, pp. 75–82.
- [24] Y. Wang, X. Lin, and M. Pedram, "A nested two stage game-based optimization framework in mobile cloud computing system," in *Service Oriented System Engineering (SOSE), 2013 IEEE 7th International Symposium on*, Mar. 2013, pp. 494–502.
- [25] V. Cardellini, V. De Nitto Personé, V. Di Valerio, F. Facchinei, V. Grassi, F. Lo Presti, and V. Piccialli, "A game-theoretic approach to computation offloading in mobile cloud computing," *Mathematical Programming*, pp. 1–29, 2015.
- [26] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Transactions on Networking*, to appear.
- [27] E. Meskar, T. D. Todd, D. Zhao, and G. Karakostas, "Energy efficient offloading for competing users on a shared communication channel," in *Proc. of IEEE ICC*, Jun. 2015, pp. 3192–3197.
- [28] X. Ma, C. Lin, X. Xiang, and C. Chen, "Game-theoretic analysis of computation offloading for cloudlet-based mobile cloud computing," in *Proc. of ACM MSWiM*, 2015, pp. 271–278.
- [29] "Congestion games with player-specific payoff functions," *Games and Economic Behavior*, vol. 13, no. 1, pp. 111 – 124, 1996.